Comments on the Summary of Statistical Analysis Package

This document describes EPA comments on the statistical methods used by the Navy at North Pier Parcel of the Hunters Point Site. In this document, comments are inserted in blue following the text described in the Navy's document, "Summary of Statistical Analysis Package."

Data were collected in many stages but this analysis only used the final status survey (FSS) and systematic survey data PRE (includes Sys_1 and Sys_2). Data were identified by type of laboratory reporting the result: Onsite or Offsite. Separate and parallel analyses were conducted on six stages of the data defined to include FSS/OnSite, FSS/Offsite, FSS/On&Offsite, PRE/Onsite, PRE/Offsite, and PRE/On&Offsite data. Each analysis consisted of the following steps.

1. This analysis described above does not evaluate FSS stage data versus PRE-stage data and will not correctly identify anomalies (representing anomalous activities) from PRE-stage to FSS stage.

2. It is highly likely that anomalous activities occurred during Sys1 and Sys 2 stages which can be identified by evaluating data (supplemented with effective graphical displays) collected during sampling phases. Combining Sys-1 and Sys-2 data as PRE-stage data will mask anomalous activities potentially present in Sys-1 and/or Sys-2 data sets (higher number of false negatives). On the other hand – the approach may identify the entire PRE (Sys-1 and Sys-2) sampling phase as anomalous (higher number of false positives) whereas only one of the sampling phase (Sys 1 or Sys 2) might be anomalous.

3. It is highly likely, that some anomalous activity took place in some intermediate sampling phase – which can be identified only by simultaneously evaluating data from all sampling phases.

1. PRELIMINARY DATA DISPLAY AND ANALYSIS

Input Dataset: Data reported in any one of the six stages defined above for a single nuclide in samples $j = 1, \dots, n_i$ from survey unit $i = 1, \dots, N$ included the specified Parcel or sub-Parcel are denoted by $D = \{d_{i,j}\} = \{D_1, D_2, \dots, D_N\}$ where $D_i = \{d_{i,1}, d_{i,2}, \dots, d_{i,n_i}\}$ are the data reported for survey unit i.

All results are identified by date. In several of the parcels the data are identified by location coordinates. Where available, the location data are used to create posting plots for each survey unit and for all survey units combined. These plots are used to identify survey unit assignment errors in the data.

How? Illustration?

The current analysis package shows the data distribution for the survey unit as a histogram, CDF, and lognormal probability plot; tables of mean value, standard deviation, coefficient of variation and precision by unit and by sampling day; a drop-line plot of the data sorted by date; time series plot for all sampling days and for days with more than four samples; a time series plot of the precision by day; and concludes with a scatter plot of daily precision versus sample size.

4. The usefulness of lognormal probability plots and histograms to identify anomalies is not clear, especially when nonparametric methods (e.g., K-S test listed below) are used. Since the objective is to identify anomalous activities, graphical displays which are better suited to identify anomalies such as box plots and quantile-quantile plots should be used.

2. NONPARAMETRIC STATISTICAL TESTS TO IDENTIFY ANOMALOUS DATA

NOTE: The procedure described in this section is applied first to compare each survey unit to the data from all other survey units. The procedure is then repeated to compare data for each survey day to data from all other survey days. This is indicated by using the phrase units/days.

5. As described below, the K-S has been used to identify anomalies. The K-S test is used to compare distributions of two data sets and not to identify anomalies. The approach described in the above NOTE implicitly assumes that data from the remaining survey units or dates (except for the one which is being compared) comes from the same population – which is not feasible to justify. It is highly likely that different survey units follow different distributions and comparing one survey unit (date) against the combined data from all survey units (incorrectly assuming the remaining survey units represent a single population) does not make sense to address the objective of identifying anomalous activities in a parcel (in a survey unit during a sampling phase on a specific date).

6. The approach used is tedious and requires unmanageable number of computations and comparisons. For an example, for a parcel with 25 relevant/important survey units sampled on 5 days for 15 RADs – this approach will require 25*15 survey unit comparisons and 15*5 date comparisons (total of 375+75=450) without any guarantee that the proposed approach will identify all anomalies correctly. Moreover, this approach will be used separately for the 6 sampling phases (described above) resulting in the computation (comparisons) of 450*6 = 2700 p-values and KS distances. Reviewing and comparing 2700 p-values (distances) is a tedious task without guaranteed success.

7. Which criterion was used for p-value (or KS distance) to determine different survey units/dates. What will be the critical level while using KS tests following a KW test for multiple RADs?

8. Ranking 2700 results (375 and 75 separately for each sampling phase) based upon their respective p-values may identify survey units (dates) which are different from the rest of the survey units (dates). However, this conclusion (significant difference) cannot be used to conclude that a survey unit (date) represents anomalous activities. K-S test determines differences between two data sets representing two populations. K-S test is not a test to identify anomalies.

9. Performing these evaluations separately for the 6 sampling scenarios will not take the sampling sequence (phase- wise and date-wise) into consideration – which perhaps is one of the most important criterion to identify anomalous activities that might have occurred in a survey unit during a sampling phase (date).

Three related nonparametric tests are applied to identify anomalous survey units/days during the survey stage in question:

1) the 2-sample Kolmogorov−Smirnov (KS) test;
2) the 2-sample, 2-dimensional Peacock test; and
3) the multi-sample Kruskal Wallace (KW) test.

The latter two tests are extensions of the KS test to multiple dimensions and multiple samples, respectively.

- K-S test determines differences between two data sets representing two populations. K-S test is not a test to identify anomalies.

10. As mentioned earlier – these tests are used to compare data from 2 or more populations and not to find anomalies. Details are presented in comments 5 -9 above.

KW Test. The KW test is applied first to determine if there are significant differences among the survey units/days. Usually, answer is yes. If the answer is yes then the K-S test is applied iteratively as described below to determine which survey unit/days look unusual when compared with the others.

11. As stated in the above paragraph, the KW test is used to compare several (>2) groups. The use of this test makes the approach more tedious resulting in many more comparisons. In our example of 10 RADs, 10 additional KW tests will be required.

KS Test. Each survey unit/day i is compared with the combined data from all other survey units/days using the 2-sample KS test. The KS test generates a p-value for the distribution in each survey unit/day indicating the probability that a difference in distributions this extreme could have occurred by chance.

$$pvalue_i = ks.test(x, y) \; where \; x = D_i \; and \; y = \{D_{j \neq i}\}$$

Let p_i denote the KS p-value obtained for survey unit/day i. Using the Bonferroni correction, the survey unit/day is flagged for further attention when $p_i < \alpha/N$ for the pre-specified value of α. Histograms and KS plots are generated in the analysis package pdf file for visual comparisons of the flagged survey unit/day distributions. After all nuclides are processed, the flagged survey units/days for all nuclides are listed and reported on one Excel spreadsheet.

How will one adjust for p-value associated with the KW tests for multiple ROCs?

The spreadsheets include the p-value for each KS test, which has been applied recently as a measure of similarity of the two distributions[1]. A high p value indicates similar distributions, while a very low p value indicates a significant difference. To identify data sets which are different than the others, a monotonic transformation of the p-value similarity measure is used to define a positive valued distance measure[2] for each survey unit/day:

$Distance_i = -\ln(p_i)$.

The survey units/days are are ranked in order of decreasing distance to expedite further investigation of this stage of survey in the parcel. This provides the same ranking as would be obtained by using the p-values directly, only in reverse order.

12. A read through the above narration of the KS test reveals that the KS test is meant for comparing data from 2 populations (e.g., such a placebo group versus treatment group). In statistical literature, this test is not used to identify observations representing anomalous activities. Also, the assumptions used to perform this test are not satisfied (by the combined survey unit data). Additional comments for KS test are provided in comments 4 through 9.

Peacock Test. The Peacock test[3] is applied to examine the equilibrium status of selected nuclides that are related by decay chain. Each survey unit i is compared with the combined data from all other survey units using Peacock's 2-dimensional analog of the 1-dimension KS test. The following four radionuclide pairs are examined: Ra226/Pb214, Ra226/Bi214, Th232/Pb212, and Th232/Bi212. The iterative procedure is very similar to that used for the KS test, with each survey unit compared iteratively to all others.

$$pvalue_i = Peacock.test(x,y) \ where \ x = D_i \ and \ y = \{D_{j \neq i}\}$$

The only difference here is that the variables x and y are matrices of two columns of data, one column for each of the two related nuclides. The survey unit is flagged for further attention due to unusual equilibrium conditions when $p_i < \alpha$ for the pre-specified value of α. The survey units flagged by the Peacock test for each nuclide recorded on a spreadsheet. Scatter plots comparing the two bivariate distributions are generated in the Peacock analysis packages for all survey units with sufficient data.

13. Comments described in 5 through 12 above related to the assumption used to compute KS distances (p-values) also apply to this Peacock test.

3. HIERARCHICAL CLUSTER ANALYSIS

HCA was performed using Ward's method[4] to group the survey units in a parcel into a small number of groups based on their KS distance matrix. The distance matrix is obtained by applying the 2-sample KS test to compare each survey unit with each other survey units. A matrix of p-values is calculated:

$$pvalue_{i,j} = ks.test(x,y) \ where \ x = D_i \ and \ y = D_j, \forall \ i,j$$

This is converted to a distance matrix using the $Distance_{i,j} = -\ln(p_{i,j})$ formula. This distance matrix is used to make the dendrogram shown in the analysis package. The purpose of this figure is to shown which other survey units should be inspected because they are similar to a survey unit found with anomalous data.

14. Did not see this approach used at the Hunters Point Site data. However, the similarity or the distance matrix generated using p-values or -log of p values is based upon the KS test. All comments associated with the KS test described in comments 5 through 12 above also apply to this approach.

4. COMPARISON OF DATA FROM ONSITE AND OFFSITE LABORATORIES

The onsite and offsite data distributions are compared using a histogram and the 2-sample KS test.

15. See comments 5 through 12.

5. SEARCH FOR DUPLICATED DATA

Each survey unit/day is compared with each other unit/day to search for duplicated. The number of matched sample values is reported when there are more than two matched results.

6. BENFORD TEST

The Benford test[5] is applied for the first and second digits. The frequency of occurrence is compared to the expected distributions.

General Comments: Methods described here (KS test, KW test, Peacock test) are intended for comparison purposes and not to identify observations representing anomalous activities (objective of these evaluations). The approach is quite tedious requiring many comparisons (based upon untenable assumptions) without any guarantee of correct identification of anomalous activities. The problem of identifying anomalous activities which might have taken place at the Hunters Point site is quite complex – and effective methods tailored made to identify patterns in data sets (supplemented with effective and useful graphical displays) should be used. The multivariate statistical methods based upon principal component analysis (PCA) are commonly used to identify of anomalies and patterns presents in complex data sets.

[1] See for example: Shaus A. and Turkel E. "Writer Identification in Modern and Historical Documents via Binary Pixel Patterns, Kolmogorov–Smirnov Test and Fisher's Method," Journal of Imaging Science and Technology 61(1):104041-104049, January 2017;

Griechisch E., Malik M., and Liwicki M. "Online signature verification based on Kolmogorov–Smirnov distribution distance", Proc. 14th Int'l. Conference on Frontiers in Handwriting Recognition, ICFHR, IEEE; Piscataway, NJ, 2014;

Jinwook S., et.al. Interactively optimizing signal-to-noise ratios in expression profiling project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays," Bioinformatics, Vol. 20, 2004.

[2] The R program *ks.test* may return a value of 0 for the p-value due to underflow when the distributions are very different. To avoid numerical difficulties with logarithms, a minimum p-value of e^{-100} is assigned when underflow occurs. This limits the measured distance to values that are between 0 and 100.

[3] Peacock J. "Two-dimensional goodness-of-fit testing in astronomy," Monthly Notices of the Royal Astronomical Society, Vol. 202, 1983.

[4] Murtagh F. and Legendre P. "Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion," Journal of Classification, 31(3), 2014.

[5] Benford F. "The Law of Anomalous Numbers," Proceedings of the American Philosophical Society Vol. 78, 1938.